Is Unequal Weighting Significantly Better than Equal Weighting for Multi-Model Forecasting?

Timothy DelSole (COLA,GMU)
Xiaosong Yang (GFDL)
Michael Tippett (IRI)
Huug van den Dool (CPC)
Liwei Jia (COLA)

October 5, 2011

Multi-model Combination of Forecasts

A linear multi-model combination is

$$y(t) = x_1(t)\beta_1 + x_2(t)\beta_2 + \cdots + x_M(t)\beta_M + \mu + \epsilon(t)$$

$y(t)$: predictand

$x_m(t)$: prediction by model m

β_m: model weight for model m

Potential Strategies for Specifying Weights

- **Linear Regression** "Super-ensemble" (Krishnamurti et al. 1999)
  - overfitting becomes a problem for large number of models M
  - weights vary substantially on short space scales
- **Ridge regression** (Peña and van den Dool 2008)
- **Multi-Model Mean** ($\beta_m = 1/M$)
- **Bayesian** (Rajagopalan et al. 2002)
  - weighting coefficients become noisy as more models included
  - neighboring grid points have very different coefficients
- **Bayesian** (DelSole 2007)
  - Nested cross validation could not beat multi-model average

Objective

Many studies show that the multi-model mean ($\beta_m = 1/M$) gives the best, or close to the best, forecast.

Multi-model mean is a special case of equal weights:

$$\beta_1 = \beta_2 = \cdots = \beta_M = \alpha/M$$

We want to test whether a multi-model combination based on unequal weights has *significantly* smaller errors than a combination based on equal weights.

Test Hypothesis of Equal Weights

$$y(t) = x_1(t)\beta_1 + x_2(t)\beta_2 + \cdots + x_M(t)\beta_M + \mu + \epsilon(t)$$

$$H_{SMMM}: \quad \beta_1 = \beta_2 = \cdots = \beta_M = \alpha/M$$

where "SMMM" stands for "scaled multi-model mean."

The statistic for testing this hypothesis is

$$F = \frac{SSE_{SMMM} - SSE_{GLM}}{SSE_{GLM}} \frac{N - M - 1}{M - 1}$$

SSE_{SMMM}: sum square error of regression model under H_{SMMM}

SSE_{GLM}: sum square error of model with least squares weights

Large F value favors a rejection of the hypothesis.

Rejection of the Hypothesis of Equal Weights

The hypothesis is

$$H_{SMMM}: \quad \beta_1 = \beta_2 = \cdots = \beta_M = \alpha/M$$

All that is required to reject H_{SMMM} is

$$\beta_i \neq \beta_j \quad \text{for at least one } i \neq j$$

This could happen in a variety of ways:

- one model has no skill ($\beta_m = 0$ for some m).
- some subset of models have no skill.

How Much Smaller Variance Does GLM Need to Explain to Reject Hypothesis of Equal Weights ?

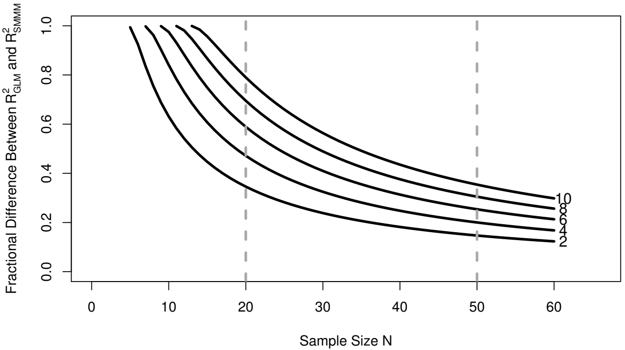
R^2_{GLM}: Fraction of variance explained by GLM.
R^2_{SMMM}: Fraction of variance explained by SMMM.

A relative measure of the difference in variances is:

$$\delta = \frac{R^2_{GLM} - R^2_{SMMM}}{1 - R^2_{SMMM}}.$$

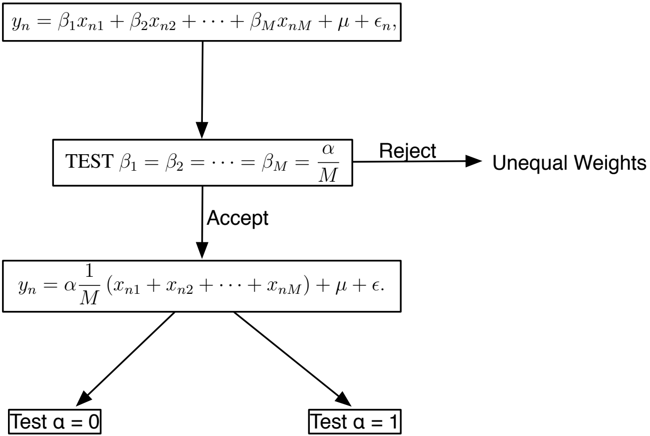
$$F = \frac{\delta}{1-\delta}\frac{N-M-1}{M-1}$$

δ Values Required to Satisfy 5% Significance Test



Different curves corresponding to different number of models (M).

Schematic of the Proposed Decision Procedure



	Accept $\alpha = 1$	Reject $\alpha = 1$
Accept $\alpha = 0$	Ambiguous	Clim
Reject $\alpha = 0$	MMM	SMMM

Test Hypothesis that Weights Vanish Simultaneously

$$y(t) = \frac{\alpha}{M}\left(x_1(t) + x_2(t) + \cdots + x_M(t)\right) + \mu + \epsilon(t)$$

$$H_{CLIM} : \alpha = 0$$

where “CLIM” stands for “climatology.”

The statistic for testing this hypothesis is

$$F = \frac{SSE_{CLIM} - SSE_{SMMM}}{SSE_{SMMM}} \frac{N-2}{1}$$

SSE_{CLIM}: sum square error of regression model under H_{CLIM}

Rejection of the Hypothesis H_{CLIM}

All that is required to reject H_{CLIM} is

$$\beta_i \neq 0 \quad \text{for at least one } i$$

This could happen in a variety of ways:

- only one model has skill ($\beta_m \neq 0$ for some m).
- all models should be equally weighted ($\alpha = 1$).

Test Hypothesis that All Weights Equal 1/M

$$y(t) = \frac{\alpha}{M}\left(x_1(t) + x_2(t) + \cdots + x_M(t)\right) + \mu + \epsilon(t)$$

$$H_{MMM} : \alpha = 1$$

where "MMM" stands for "multi-model mean."

The statistic for testing this hypothesis is

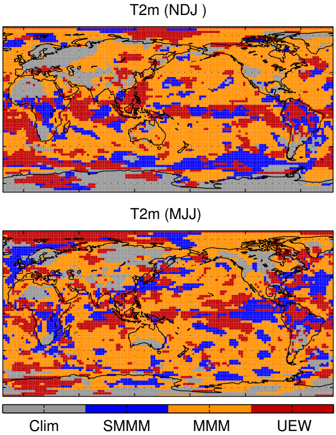
$$F = \frac{SSE_{MMM} - SSE_{SMMM}}{SSE_{SMMM}} \frac{N-2}{1}$$

SSE_{MMM}: sum square error of regression model under H_{MMM}

Application to Seasonal Hindcasts

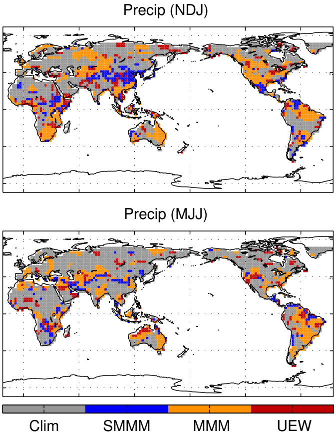
- ENSEMBLES data set (Weisheimer et al., 2009)
  - UK Met
  - Météo France
  - ECMWF
  - Leibniz Institute of Marine Sciences at Kiel University
  - Euro-mediterranean Centre for Climate Change in Bologna
- 9 member ensemble
- consider only hindcasts initialized 1 May and 1 November
- 46 year period 1960-2005
- NDJ and MJJ mean 2m temperature and precipitation
- 2m temperature verified against NCEP/NCAR reanalysis
- precipitation verified against NCEP/CPC (Chen et al. 2002)

Selected Strategies for 2m Temperature



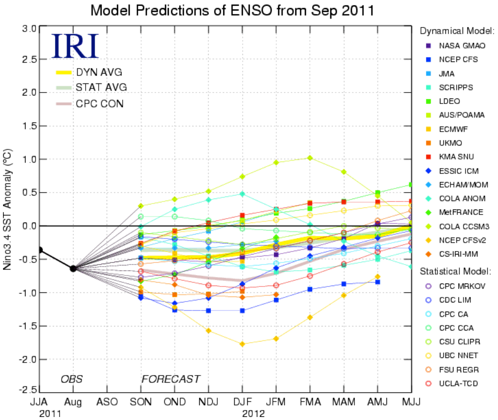
- Equal weights can not be rejected over 3/4 of the globe.
- Multi-model mean is the dominant choice.

Selected Strategies for Precipitation



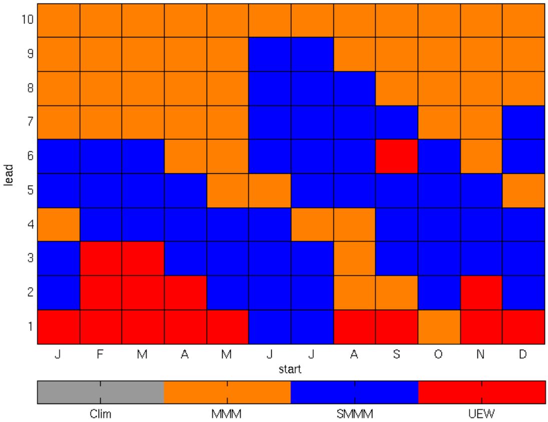
- Equal weights can not be rejected over 90% of the land.
- Vanishing weights is the dominant choice.

IRI Plume



- Apply tests to hindcasts of 3-month average NINO3.4
- 28-29 years of data (1982-2010).
- 5-15 ensemble members, depending on lead
- Test for each initial month and lead.

Selected Strategies for IRI Plume



- For short lead time, unequal weights is the dominant choice.

Summary

- We proposed statistical test for whether a multi-model combination with unequal weights has significantly smaller errors than a combination with equal weights.
- If hypothesis of equal weights is rejected, this test gives no information about the best strategy for unequal weighting.
- Equal weights could not be rejected over three-quarters of the globe for T2m, and 90% for land precipitation.
- For equal weighting, multi-model mean was the dominant choice for T2m, and vanishing weights for precipitation.
- For IRI plume, unequal weighting was selected mostly for short leads, presumably because models are distinguishable at high skill.
- For IRI plume, climatology is not selected.